

SchistoDB: an updated genome resource for the three key schistosomes of humans

Adhemar Zerlotini^{1,2}, Eric R. G. R. Aguiar¹, Fudong Yu³, Huayong Xu⁴, Yixue Li^{3,4}, Neil D. Young⁵, Robin B. Gasser⁵, Anna V. Protasio⁶, Matthew Berriman⁶, David S. Roos⁷, Jessica C. Kissinger⁸ and Guilherme Oliveira^{1,*}

¹Centro de Excelência em Bioinformática, National Institute for Science and Technology in Tropical Diseases FIOCRUZ-Minas, Belo Horizonte, MG, 30190-002, ²Laboratório Multiusuário de Bioinformática, Embrapa Informática Agropecuária, Campinas, SP, 13083-970, Brazil, ³Shanghai Center for Bioinformation Technology, Shanghai, China, ⁴School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China, ⁵Faculty of Veterinary Science, University of Melbourne, Parkville, Victoria, 3010, Australia, ⁶Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK, ⁷Department of Biology, University of Pennsylvania, PA and ⁸Center for Tropical and Emerging Global Diseases, University of Georgia, GA, 30602-7399, USA

Received September 20, 2012; Accepted October 17, 2012

ABSTRACT

The new release of SchistoDB (<http://SchistoDB.net>) provides a rich resource of genomic data for key blood flukes (genus *Schistosoma*) which cause disease in hundreds of millions of people worldwide. SchistoDB integrates whole-genome sequence and annotation of three species of the genus and provides enhanced bioinformatics analyses and data-mining tools. A simple, yet comprehensive web interface provided through the Strategies Web Development Kit is available for the mining and visualization of the data. Genomic scale data can be queried based on BLAST searches, annotation keywords and gene ID searches, gene ontology terms, sequence motifs, protein characteristics and phylogenetic relationships. Search strategies can be saved within a user's profile for future retrieval and may also be shared with other researchers using a unique web address.

INTRODUCTION

In the last few years, major advances have been made in the sequencing of genomes of the blood flukes of the genus *Schistosoma* which causes chronic diseases in ~200 million people globally (1). Thus far, the focus has been on *Schistosoma haematobium*, *Schistosoma japonicum* and *Schistosoma mansoni*. Recently, *S. haematobium* had its genome sequenced (2) and published as well as a new and improved assembly of the first-sequenced schistosome genome, *S. mansoni*, was made available (3). The initial

version of SchistoDB (<http://SchistoDB.net>) (4) was implemented to provide easy access and visualization of the *S. mansoni* genome and features (5), integrated to other data types such as expressed sequence tags (ESTs), proteins and metabolic pathways. In order to interrogate datasets from multiple genomes in a comprehensive manner, a new version of SchistoDB was developed as a resource for genomic data across the genus *Schistosoma*.

This was done in partnership with the NIAID-funded Eukaryotic Pathogen Bioinformatics Resource Center (<http://EuPathDB.org>) (6). This enhanced database uses the same database structural framework and uses the graphical Strategies Web Development Kit (WDK) search interface (7). It also provides a data-mining interface for the comparative and functional genomic data of three species of schistosomes and an integrated query system as part of the WDK and Genomics Unified Schema database structure. SchistoDB differs from other resources such as the Wellcome Trust Sanger Institute GeneDB (<http://GeneDB.org>) which has complementary data query or visualization tools but do not have the data-mining capabilities and broad cross-species comparisons which are possible using the WDK search interface. Data are currently obtained directly from providers at sequencing centers, GenBank and associated functional data repositories.

CONTENTS OF THE CURRENT RELEASE

The current release of SchistoDB contains the latest release of the genome sequence and annotation from *S. haematobium* (2), *S. japonicum* (8) and *S. mansoni* (3) provided by the Faculty of Veterinary Science at the

*To whom correspondence should be addressed. Tel: +55 31 3337 3649; Fax: +55 31 3295 3115; Email: oliveira@ceb.io.org

Table 1. Species in the current release of SchistoDB

Species	Strain	Genome size (Mb)	Number of genes	Reference
<i>Schistosoma haematobium</i>	Egypt	385	13 073	(2)
<i>Schistosoma japonicum</i>	Anhui	397	12 669	(8)
<i>Schistosoma mansoni</i>	Puerto Rico	380	12 871	(3)



Figure 1. Screen shot of the SchistoDB homepage. (A) The banner section present on all SchistoDB webpages provides links to registration, login, and contact us forms, ID and text searches, information and help, and all available searches and tools. (B) The side panel on the homepage provides expandable tabs that reveal links to a data summary table, community news items, links to community resources, web tutorials and additional information and help links. (C) The central portion of the home page includes all available searches and tools—gene-specific searches, searches against other data types such as genomic sequence, open reading frames and ESTs and tools, such as BLAST, sequence retrieval and the genome browser.

University of Melbourne, the Chinese National Human Genome Center and the Wellcome Trust Sanger Institute, respectively. More information about these genomes is available in Table 1.

Genomes and annotation in SchistoDB are processed through the same analysis pipeline, which provides additional data, including InterPro domains (9), gene ontology term association (10), signal peptide predictions (11), transmembrane domain predictions, open reading frame predictions, BLAST against the non-redundant genome database at the National Center for Bioinformatics, orthology prediction based on OrthoMCL (12) and synteny prediction. Pipeline details are available at <http://schistodb.net/schisto/showXmlDataContent.do?name=XmlQuestions.Methods>.

HOW TO USE SCHISTODB

The home page

The SchistoDB home page is virtually the same as all EuPathDB pages, with differences only in color, logo and data content. A visitor to these sites will first notice the home page layout, which has been designed to provide

the user with convenient and immediate access to data and tools. The home page is divided into three main sections (Figure 1). (i) A top banner section, providing quick access back to the home page, gene ID and text searches, (ii) 'contact' and login/registration links and (iii) mouse over menus (Figure 1A). The information and help menus on the left (Figure 1B). The central section provides links to all searches (Figure 1C), and links to tools, such as BLAST (13), the sequence retrieval tool and the Generic Model Organism Database genome browser (14). Creating an account and logging in allows search strategies to be saved and shared, and gene associated annotation comments to be created which are linked to the author.

Building a search strategy

The search strategy system allows users to filter the results list based on a combined set of criteria and also add or sort columns. After running the first search (Figure 2), a user might elect to add other filtering steps. This can be achieved by sequentially adding new searches to grow the strategy horizontally. Steps in a strategy may be viewed, revised, renamed and developed further by nesting or deleted. Entire search strategies may be renamed, copied,

The screenshot displays the SchistoDB web interface. At the top, the header includes the SchistoDB logo, version 3.0 (1 Jul 12), and the EuPathDB Project affiliation. Navigation tabs include Home, New Search, My Strategies, My Basket (4), Tools, Data Summary, Downloads, Community, and My Favorites. The 'My Strategies' section shows a workflow diagram with four steps: Step 1 (Signal Pep, 4192 Genes), Step 2 (Transmb Dom, 4710 Genes), Step 3 (Egg EST, 5216 Genes), and Step 4 (Orthologs, 5059 Genes). An 'Add Step' button is visible. A 'Secreted Egg Proteins' popup is open, showing a URL to copy and paste: <http://schistodb.net/schisto/im.do?s=ea7002f6e5b2996d>. Below the workflow, a table titled 'Secreted Egg Proteins - step 4 - 5059 Genes' is shown, with options to 'Add 5059 Genes to Basket' or 'Download 5059 Genes'. The table has columns for Gene, Organism, Genomic Location, and Product Description. The footer includes the SchistoDB 3.0 logo, version 3.0 (1 Jul 12), and the EuPathDB Project Team, along with a contact link.

Gene	Organism	Genomic Location	Product Description
Sjp_0010400	<i>S. japonicum</i>	SJC_S000039: 99,106 - 101,313 (+)	unspecified product
Sjp_0018050	<i>S. japonicum</i>	SJC_S000087: 563,316 - 565,488 (+)	unspecified product
Sjp_0044680	<i>S. japonicum</i>	SJC_S000376: 75,892 - 80,845 (+)	unspecified product
Sjp_0066310	<i>S. japonicum</i>	SJC_S000836: 144,761 - 145,739 (+)	unspecified product
Smp_049550	<i>S. mansoni</i>	Schisto_mansoni.Chr_1: 49,952,027 - 49,954,197 (+)	heat shock protein 70 (hsp70), putative
Smp_106930.1	<i>S. mansoni</i>	Schisto_mansoni.Chr_1: 47,704,341 - 47,707,076 (+)	heat shock protein 70, putative
Smp_106930.2	<i>S. mansoni</i>	Schisto_mansoni.Chr_1: 47,704,341 - 47,707,076 (+)	heat shock protein 70, putative
Smp_186050	<i>S. mansoni</i>	Schisto_mansoni.Chr_1: 47,704,538 - 47,706,451 (+)	heat shock protein, putative
Sha_100375	<i>S. haematobium</i> Egypt	scaffold8: 578,901 - 607,312 (+)	Collagen alpha-1(XXVII) chain, putative
Sha_101247	<i>S. haematobium</i> Egypt	scaffold267: 415,858 - 474,039 (-)	Collagen alpha-1(IV) chain, putative
Sha_101424	<i>S. haematobium</i> Egypt	scaffold191: 83,774 - 100,919 (+)	Collagen-like protein 4, putative
Sha_101757	<i>S. haematobium</i> Egypt	scaffold267: 401,340 - 426,495 (-)	Collagen alpha-1(IV) chain, putative
Sha_102615	<i>S. haematobium</i> Egypt	scaffold342: 621,131 - 644,257 (+)	Collagen alpha-1(XXVII) chain, putative
Sha_103595	<i>S. haematobium</i> Egypt	scaffold191: 52,851 - 78,098 (-)	Collagen alpha-2(II) chain, putative
Sha_104571	<i>S. haematobium</i> Egypt	scaffold401: 97,109 - 169,556 (-)	Collagen alpha-1(IV) chain, putative
Sha_105038	<i>S. haematobium</i> Egypt	scaffold191: 102,343 - 132,689 (+)	Collagen alpha-1(XXVII) chain, putative
Sha_106530	<i>S. haematobium</i> Egypt	scaffold315: 17,556 - 62,008 (-)	Collagen alpha-1(XXVII) chain, putative
Sha_107194	<i>S. haematobium</i> Egypt	scaffold401: 229,106 - 270,401 (+)	Collagen alpha-1(XXVII) chain, putative
Sha_107460	<i>S. haematobium</i> Egypt	scaffold509: 582,345 - 624,044 (-)	Collagen-like protein 2, putative
Sjp_0000750	<i>S. japonicum</i>	SJC_S000001: 961,569 - 968,429 (-)	unspecified product

Figure 2. Screen shot of a search strategy with the first step being the results of running the search shown in Figure 1C (inset). The 'Add Step' button reveals a popup with all available searches in SchistoDB. The results of any search are displayed in a dynamic table that allows removing, adding and moving columns, downloading results and adding results to the basket.

saved and shared with a unique strategy URL or deleted. An example of a complex multi-step search strategy can be seen in Figure 2. Using this strategy, for example, all secretory genes expressed in eggs are identified. This is achieved by finding all genes with predicted secretory signal peptides and/or transmembrane domains (Steps 1 and 2, Figure 2), and that have egg EST libraries mapping evidence (Step 3, Figure 2). As a final step, a transformation is applied on the results to identify all *Schistosoma* orthologs of the results in Step 3 (Step 4, Figure 2) since there are only egg EST libraries for *S. mansoni*. Several options can be applied to a whole strategy including renaming, copying, saving, deleting and sharing. The latter allows users to email colleagues a unique URL of a strategy of interest, which enables the receiver to open and modify the strategy in their own workspace (for example, the strategy in Figure 2 can be accessed here: <http://schistodb.net/schisto/im.do?s=ea7002f6e5b2996d>).

Additional features

There is a range of features that allows users to bookmark their favorite genes for quick future access; add genes to a basket in order to combine such gene set in later search;

add arbitrary weights to steps to obtain a ranked list or write comments to genes to improve its annotation. Data in SchistoDB are conveniently available for bulk download from the 'Data Files' section accessible from the 'Downloads' menu item in the gray tool bar (Figure 1A). Data files are in folders organized by database release version number and species. The sequence retrieval tool, accessible from the tools section (Figure 1C), allows users to specify exact coordinates or lists of genes or proteins to be downloaded.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the genome sequencing centers: the Wellcome Trust Sanger Institute, the Faculty of Veterinary Science at the University of Melbourne and the Chinese National Human Genome Center for providing the genome assembly and annotation. Without their generous pre-publication contribution developing this integrated database resource would not have been possible. Special thanks to the EupathDB team, which provided skills, expertise and the technology to accomplish this work.

FUNDING

National Institutes of Health (NIH)—Fogarty International Center [TW007012-03]; The Burroughs Wellcome Fund (BWF); CNPq [573839/2008-5]; FAPEMIG [CBB-1181/08485/2009, REDE-56/11] and EC FP7 [241865]. Funding for open access charge: NIH.

Conflict of interest statement. None declared.

REFERENCES

1. Rollinson, D.A. (2009) Wake up call for urinary schistosomiasis: reconciling research effort with public health importance. *Parasitology*, **136**, 1593–1610.
2. Young, N.D., Jex, A.R., Li, B., Liu, S., Yang, L., Xiong, Z., Li, Y., Cantacessi, C., Hall, R.S., Xu, X. *et al.* (2012) Whole-genome sequence of *Schistosoma haematobium*. *Nat. Genet.*, **44**, 221–225.
3. Protasio, A.V., Tsai, I.J., Babbage, A., Nichol, S., Hunt, M., Aslett, M.A., De Silva, N., Velarde, G.S., Anderson, T.J.C., Clark, R.C. *et al.* (2012) A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. *PLoS Negl. Trop. Dis.*, **6**, e1455.
4. Zerlotini, A., Heiges, M., Wang, H., Moraes, R.L.V., Dominitini, A.J., Ruiz, J.C., Kissinger, J.C. and Oliveira, G. (2009) SchistoDB: a *Schistosoma mansoni* genome resource. *Nucleic Acids Res.*, **37**, D579–D582.
5. Berriman, M., Haas, B.J., LoVerde, P.T., Wilson, R.A., Dillon, G.P., Cerqueira, G.C., Mashiyama, S.T., Al-Lazikani, B., Andrade, L.F., Ashton, P.D. *et al.* (2009) The genome of the blood fluke *Schistosoma mansoni*. *Nature*, **460**, 352–358.
6. Aurrecochea, C., Brestelli, J., Brunk, B.P., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O.S., Heiges, M. *et al.* (2010) EuPathDB: a portal to eukaryotic pathogen databases. *Nucleic Acids Res.*, **38**, D415–D419.
7. Fischer, S., Aurrecochea, C., Brunk, B.P., Gao, X., Harb, O.S., Kraemer, E.T., Pennington, C., Treatman, C., Kissinger, J.C., Roos, D.S. *et al.* (2011) The strategies WDK: a graphical search interface and web development kit for functional genomics databases. *Database*, 2011, bar027.
8. Liu, F., Zhou, Y., Wang, Z.-Q., Lu, G., Zheng, H., Brindley, P.J., McManus, D.P., Blair, D., Zhang, Q.-hua, Zhong, Y. *et al.* (2009) The *Schistosoma japonicum* genome reveals features of host–parasite interplay. *Nature*, **460**, 345–351.
9. Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. and Lopez, R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.
10. Gene Ontology Consortium. (2012) The Gene Ontology: enhancements for 2011. *Nucleic Acids Res.*, **40**, D559–D564.
11. Bendtsen, J.D., Nielsen, H., von Heijne, G. and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.
12. Chen, F., Mackey, A.J., Stoeckert, C.J. and Roos, D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.
13. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
14. Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.